

CAI
BS 12
69C13



~~Statistics Bureau 87
General publications,
-49, An introduction to the
socio-economic information
system (CANSIM)~~

CANSIM

CANSIM:

AN INTRODUCTION TO THE CANADIAN SOCIO-ECONOMIC INFORMATION MANAGEMENT SYSTEM



Canada
DOMINION BUREAU OF STATISTICS

The contents of this document may be used freely but DBS should be credited when republishing all or any part of it.

CA1
BS 12
-69C13

DOMINION BUREAU OF STATISTICS
National Accounts, Production and Productivity Division
General Time Series Section

INTRODUCTION TO THE CANADIAN SOCIO-ECONOMIC
INFORMATION MANAGEMENT SYSTEM (CANSIM)



Published by Authority of
The Minister of Trade and Commerce

Special Reprint from the Canadian Statistical Review, March, 1969




INTRODUCTION TO THE CANADIAN SOCIO-ECONOMIC INFORMATION MANAGEMENT SYSTEM (CANSIM)

Mary Lennox and T.J. Vander Noot*

TABLE OF CONTENTS

	Page
Introduction	v
Acknowledgements	v
Definitions	v
Description of CANSIM	v
General	v
The Structure of the Base and the Data Entry Sub-system	vii
Hardware	ix
Programming Languages	ix
CANSIM Services: Availabilities and Costs	
Retrieval Requests	ix
Cost of Services	ix
Contents of the Data Base	ix
Responsibility for Data in the Data Base	x
Documentation Available from DBS	x
The DATABANK — MASSAGER System	x

* Miss Lennox is Chief of the General Time Series Section which has responsibility for the CANSIM operation and the data base. Dr. T.J. Vander Noot, formerly with the Economic Council of Canada, is now Associate Director General, Operations and Systems Development Branch, DBS.



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761116329103>

INTRODUCTION TO THE CANADIAN SOCIO-ECONOMIC INFORMATION MANAGEMENT SYSTEM (CANSIM)

Introduction

The creation of a fully computerized national data base containing time series at all levels of aggregation is now a distinct possibility with development by the Dominion Bureau of Statistics of the Canadian Socio-economic Information Management System (CANSIM).

CANSIM is a long-term project, begun in 1967 and building on earlier work of the Economic Council and DBS. For a number of reasons, including costliness and the scarcity of skilled programming resources, the project is being developed in additive phases, each complete and each representing a possible terminal point. On completion of each phase there is an administrative review, a re-evaluation of the project and allocation of resources for the next phase. This phasing of development also permits continuation of the user-designer dialogue which originally underscored the value of a data bank and brought DBS into this project.

The first stage is now operational. This provides for storage and maintenance of the data base in direct access memory, and efficient and frequent handling of a large number of series. It also provides information and carries out house-keeping tasks required for the operation and administration of a service to users.

When fully developed, the system will provide machine readable data, from a large base, possibly encompassing the total output of published statistics from DBS and other data sources. Manipulative capability, to be incorporated into the system, will allow users to select a series, and any one of a variety of standard statistical techniques to be applied to that series.

As a national data base CANSIM offers the potential to increase productivity of all economic statisticians and economists, both public and private. A reduction in research costs is also forecast because the largest expense in much social research today is the preparation of data to the computer.

Bonuses within DBS will include improved timeliness, efficiency, and lower costs in publishing information. It should also make available much publishable information not presently published by conventional methods.

Acknowledgements

DBS acknowledges with gratitude the continuing support and co-operation of the Economic Council of Canada. Council staff members who were active in the project were Dr. T.J. Vander Noot, the designer of CANSIM who has recently joined DBS, Mr. Aurele Leduc and Miss Denise Roussin.

The Bank of Canada and the Department of Finance participated in the project as members of an Interdepartmental Users Group. In addition, the

Bank is making available two computer programs, one a manipulative package, which may be used in conjunction with CANSIM.

Within DBS, Mr. R. Tharp and Mrs. R. Webster made a significant contribution as members of the team carrying out the system development. In the General Time Series Section, Mr. T. Tanaka, Head of the Publication and CANSIM Users' Service, and Miss Eileen Routliffe worked with a small staff in volume testing and implementation of the system. In co-operation with subject matter staffs they validated information in the data base and instituted procedures required for timely entry of current data.

Definitions

The following definitions of terms used in this article are provided to assist readers not familiar with computers and data banks: **Data** refers generally to time series, defined as observations through time which have some common characteristics, such as the number of births recorded annually in Canada. (Information from individual records such as the Census are excluded from this discussion.) **Data file** refers to a set of statistical observations transformed into machine-readable form, for example, punched cards or punched card images on magnetic tape. A **data base** is a set of related data files. A **data bank** is a system by which the data in the data base can be easily and inexpensively stored, maintained, retrieved and manipulated. Under this definition, CANSIM qualifies as a data bank. An **information management system** is a software tool for organizing, processing and presenting information. **Software** means the programs necessary to operate or use the computer. As computers grow increasingly sophisticated, of necessity the sophistication and importance of the software also increases. **Hardware** refers to the computers and related equipment. **Direct access memory** permits records to be stored sequentially or randomly and enables record retrieval without sequential searching through a file. **COBOL** stands for Common Business Oriented Language. **DATABANK** and **MASSAGER** are the names given to two computer programs for data storage, and for retrieval and manipulation of the stored data.

Description of CANSIM

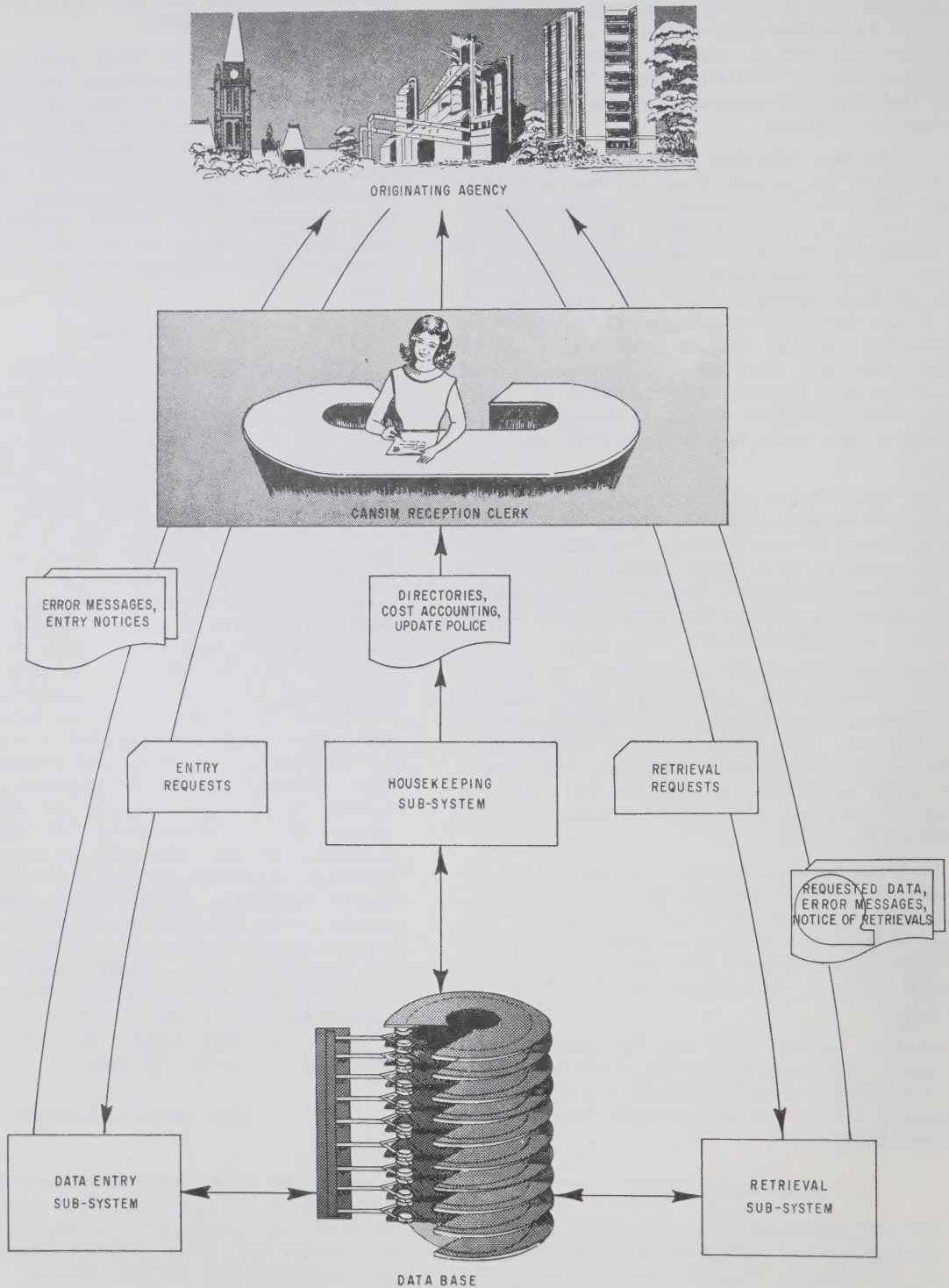
General

CANSIM is a generalized information system which combines large data files with tailor-made retrieval and manipulative packages. Operation of the system and of the CANSIM Users' Service is the responsibility of the General Time Series Section of DBS.

CANSIM provides for storage of time series in a generalized format. A data entry program creates new files, updates existing files, permits revision and correction of stored data and contains a number of built-in checks to minimize errors being introduced into the base. There is also a housekeeping

FIGURE-1

CANSIM: RELATIONSHIP OF SUB-SYSTEMS



system which re-sorts files, polices required updating of files, prepares directories of series included, and provides cost-accounting and other information required to operate and administer the system. Both systems are now in operation.

Development of the full retrieval system has been postponed to a future development stage in view of the priority given to the data entry and housekeeping programs.

As the retrieval system is modified and expanded the "CANSIM Users' Manual for Retrieval and Manipulation" will also be amended. Consistent with the central design criterion, however, users will be protected, that is, "what worked yesterday will continue to work today even though additional commands are available".

Plans include incorporation of a range of standard statistical routines, (such as the US Bureau of the Census' X-II Seasonal Adjustment program). This will allow users to select the series they wish, and the statistical routine they wish to have applied to the series.

The system as a whole, and the relationship of its component sub-systems, are illustrated in Figure 1. This shows that all action requests flow out of, and completed action requests flow into "originating agencies"—which the system recognizes by appropriate codes—through a central CANSIM Reception Clerk. Any federal government agency may store and maintain information of interest to them and may retrieve this or other information from the base. On the data entry side the largest user is DBS where the General Time Series Section is responsible for storing and maintaining information of general interest to government and public users. On the retrieval side, the CANSIM Users' Service in DBS submits retrieval action requests on behalf of non-government users.

The data entry sub-system is the only one which has the power to write on the base. The confidentiality passwords in the matrix record include a "data entry security word" without which it is impossible to enter or change information in a matrix. Other confidentiality passwords in the matrix record and in the series record protect information against unauthorized withdrawals. Agencies responsible for the data are notified via

the housekeeping system of successful and unsuccessful retrieval requests for their confidential series, and of the passwords used.

The Structure of the Base and the Data Entry Sub-system

The heart of such a system is naturally the data base, and perhaps the most crucial design element in CANSIM is that direct access memory will be used to hold the entire data base. Each time series is entered as part of a matrix of related files arranged in hierarchical fashion as illustrated in Figure 2. The matrix arrangement exploits the interdependence usual in time series to achieve efficiency and economy in file maintenance. All series in the matrix can be updated or revised at the same time, thus reducing the most costly computer operation, the accessing of the data in the base. The matrix arrangement also makes possible a greater degree of internal verification of the data entered. For instance, in Figure 2, the 03 level entries must add to the preceding 02 level and the 02 levels must add to the 01 level. Figures 3 and 4 show a population table as it actually appears in the "Canadian Statistical Review" and in the CANSIM Directory of available series. Information concerning the data is stored on a matrix record and on series records. The matrix record includes information common to all such series such as title, source notes, confidentiality passwords and an indication as to whether or not the addition check is desired as the data is entered. In the directory example, Figure 3, "crossfoot performed" indicates that the check has been made. Information stored in the series record relates only to the specific series and includes information required for arithmetic operations by computer, for printing out tables from the base, for checks on accuracy of entries, and for policing of up-dating.

An important feature of the data entry program which provides for future development of sophisticated retrievals and manipulation of data, and for the full printing of full-scale reports from the base is inclusion of descriptive attributes on each and every data point: date of reference, date of entry, publication attributes (whether the figure is preliminary, estimated, or revised), codes for retrieving footnotes from the matrix record where required, and the security or confidentiality code.

FIGURE 2:

CANSIM Hierarchical Base Structure

Level	Contents
00	Title, source, footnotes, interaction and label statements
01	Total population of Canada
02	Total male population
03	Total male population 14 to 64 years of age
03	Total male population 65 and over
02	Total female population
03	Total female population 14 to 64 years of age
03	Total female population 65 and over

FIGURE 3:

Canadian Statistical Review

Population statistics

January, 1969

Table 1: population, by province (thousands)

years and months	Canada	Nfld.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Yukon	N.W.T.
1966 June	20,015	493	109	756	617	5,781	6,961	963	955	1,463	1,874	14	29
1967 June	20,405	500	109	757	620	5,868	7,149	963	958	1,490	1,947	15	29
1966 Jan.	19,857	490	108	754	616	5,740	6,888	962	952	1,456	1,848	15	28
Apr.	19,939	492	108	755	616	5,762	6,926	963	954	1,459	1,862	14	28
June	20,015	493	109	756	617	5,781	6,961	963	955	1,463	1,874	14	29
July	20,050	494	109	756	617	5,788	6,979	963	956	1,465	1,880	14	29
Oct.	20,158	496	109	755	617	5,812	7,033	961	957	1,470	1,905	14	29
1967 Jan.	20,252	497	109	755	618	5,833	7,078	959	956	1,476	1,927	15	29
Apr.	20,334	500	109	756	619	5,854	7,115	961	955	1,483	1,938	15	29
June	20,405	500	109	757	620	5,868	7,149	963	958	1,490	1,947	15	29
July	20,441	501	109	758	621	5,873	7,167	965	958	1,493	1,952	15	29
Oct.	20,548	502	109	758	623	5,894	7,217	966	959	1,502	1,973	15	30
1968 Jan.	20,630	502	110	760	623	5,910	7,252	968	959	1,511	1,990	15	30
Apr.	20,700	505	110	760	624	5,923	7,283	969	959	1,520	2,002	15	30
June	20,744	507	110	760	624	5,927	7,306	971	960	1,526	2,007	15	31
July	20,772	508	110	760	625	5,930	7,321	972	961	1,529	2,010	15	31
Oct.	20,857	511	110	762	626	5,945	7,355	974	962	1,538	2,028	15	31

Source: Estimated population of Canada, by province (91-201), D. B. S.

FIGURE-4

SAMPLE MATRIX DIRECTORY

Matrix number 000001 NUMBER OF PERSONS IN CANADA, BY PROVINCES, BY QUARTERS SINCE 1940.

Matrix title EST. POPULATION OF CANADA BY PROV. (91-201), D.B.S.

Matrix note ESTIMATES FOR CALENDAR QUARTERLY PERIODS, FROM JULY 1, 1951. QUARTERLY DATA RELATE TO JAN.1, APR.1, JULY 1, AND OCT. 1.

Responsible Agency and Section DBS2 6002 CROSSFOOT PERFORMED

Series number	1	CANADA	01-01-40	PUBLIC
	1.1	NEWFOUNDLAND	01-01-46	PUBLIC
	1.2	PRINCE EDWARD ISLAND	01-01-47	PUBLIC
	1.3	NOVA SCOTIA	01-01-40	PUBLIC
	1.4	NEW BRUNSWICK	01-01-40	PUBLIC
	1.5	QUEBEC	01-01-40	PUBLIC
	1.6	ONTARIO	01-01-40	PUBLIC
	1.7	MANITOBA	01-01-40	PUBLIC
	1.8	SASKATCHEWAN	01-01-40	PUBLIC
	1.9	ALBERTA	01-01-40	PUBLIC
	1.10	BRITISH COLUMBIA	01-01-40	PUBLIC
	1.11	YUKON	01-01-40	PUBLIC
	1.12	NORTHWEST TERRITORIES	01-01-40	PUBLIC

Sum of LEVEL
O2 series should
equal LEVEL O1Security level of
Series:"Public" says
the series are
published or
publishable

LEVEL O2 Series

LEVEL O1 (high level) Series

Starting date of
series in the base

Hardware

CANSIM has been programmed for an IBM 360 System (Model 40 or larger). It is currently operational on a Model 65 at the Central Data Processing Service Bureau and will later be transferred to a Model 65 at DBS. The data base is held on a 2314 disc system which provides demountable direct access memory storage (necessary when a system operational at a service bureau must protect the confidentiality of data).

Programming Languages

Apart from a few small Assembler (machine language) sub-routines, all programming is in COBOL.

CANSIM Services: Availabilities and Costs

Early in the new fiscal year, DBS will offer machine-readable data from the CANSIM data base (initially some 7000 time series). A formal announcement will be made jointly by DBS and the Economic Council of Canada. Details of what retrievals are possible, and instructions for obtaining data, are contained in "CANSIM: Users' Manual for Retrieval and Manipulation". Four retrieval "commands" are available:

1. Retrieve on tape in MASSAGER format.

The tape obtained using this command is compatible with and can be used as input to the DATABANK-MASSAGER programs.

2. Retrieve on tape in PUBLICATION format.

The tape obtained using this command contains data for the series requested, plus all information necessary for publication of reports by a report generating routine.

3. Retrieve in TABLE format.

This command produces a table for those "who wish to see what is in the data base." It is a working command of interest primarily to users in DBS and other federal agencies which store and maintain data in CANSIM.

4. Retrieve on tape/cards in RE-ENTRY format.

This command, which provides for needs of the agencies using the **Data Entry Program**, can be used to supply data in card images on tape.

The manipulative capability is still so rudimentary as to be almost non-existent. Users who do not wish to write their own manipulative programs, may obtain the MASSAGER program at nominal cost from the Bank of Canada, Ottawa.

Retrieval Requests

Series may be selected individually from the total data base or users may purchase a **standard tape** in "MASSAGER format" which contains all series in the "Canadian Statistical Review". The **standard tape** will be expanded to 5000 series as quickly as time and resources permit.

There are no procedures for supplying machine-readable data to update tapes once they have been purchased. Users could update tapes in "MASSAGER format" by key punching the data for the DATABANK program (obtainable from the Bank of Canada), but a simpler procedure will be to purchase fresh tapes from the updated CANSIM base.

Requests received in DBS will normally be submitted for execution at the close of business each day, and completed jobs will be dispatched by the close of business on the following day whenever possible. The turnaround time in DBS, therefore, should not exceed 24 hours under normal circumstances. At this time there are no concrete plans for real-time operation and remote terminals although these are included in future goals.

Cost of Services

Individual series, retrieved to order of user, on user's tape, cost 15 cents per series per retrieval, with a minimum of \$25.00.

The **standard tape** containing the 2500 series in the "Canadian Statistical Review" (to be increased to 5000), in "MASSAGER format", on user's tape, costs \$100.00 per copy.

Charges have been established on the basis of quite limited evidence as to demand and will be reviewed and revised on a continuing basis. All contracts are renegotiable as of April 1, 1970.

Contents of the Data Base

CANSIM contains time series, for the most part published by DBS. The contents of the data base at April 1, 1969, are shown in Table 1.

Current information is added to the data base and is available for retrieval as soon as possible after its release by the data source. All series in the "Canadian Statistical Review" are normally available for retrieval by the 15th of the month of issue.

Expansion of the total data base will proceed as quickly as time and resources permit. The CANSIM system has been designed to be efficient and economic from 20,000 to 1.5 million series (this upper limit, imposed by currently available hardware, may disappear as new equipment is developed). Expansion could thus bring into the data base all publishable series produced in DBS as well as many produced elsewhere. Because all series in the data base are updated regularly by entry of current and revised information, the growth of the data base results in a matching growth in the current maintenance load. This is a significant factor determining the rate of growth when the maintenance is manual—that is, involving clerical preparations and key punching. The effective rate of expansion, therefore, is governed not so much by technical limitations as

by the sometimes conflicting factors of time and resources and by the speed and direction of data processing automation in DBS.

For all series, historical data are in the data base from 1946 wherever possible, and from earlier years in some instances.

Responsibility for Data in the Data Base

The introduction of CANSIM does not affect in any way relationships which may exist between users of DBS statistics and the DBS divisions responsible for their production and for their quality and confidentiality.

TABLE 1. Contents of the Data Base as of February, 1969

- (1) "Canadian Statistical Review" Catalogue No. 11-003

Some 2,500 series currently published in the Review are carried back to 1946 for most series. Includes also a small number of terminated series, and of series included for checking purposes only.

- (2) "Prices and Price Indexes" Catalogue No. 62-002

Includes 1,800 monthly and annual series with historical data as far back as 1935 when available.

- (3) "Balance of Payments" Catalogue Nos. 67-001, 67-201, 67-505

Includes 600 quarterly and annual series from Tables 1 and 2 of the "Canadian Balance of International Payments, a Compendium of Statistics from 1946".

- (4) "Agriculture Farm Finance" Catalogue No. 21-511

Includes 1,100 annual series, with data for the most part from 1926 or 1947.

- (5) "National Accounts, Income and Expenditure" Catalogue No. 13-001

In addition to series included in the "Canadian Statistical Review", there are 30 quarterly series available from 1947, and 600 annual series beginning in 1926.

- (6) "Real Domestic Product" Catalogue No. 61-506

Includes 300 quarterly and annual volume indexes (1961=100), based on the 1960 S.I.C. from 1961. These are supplementary to series published in the C.S.R.

- (7) "Industrial Production" Catalogue No. 61-506

Includes 160 annual volume indexes (1961=100) based on the 1960 S.I.C. from 1919 or earliest available year. These are supplementary to series published in the C.S.R.

Documentation Available from DBS

"CANSIM: Operational Manual for Data Entry", Catalogue No. 12-530, price \$1.00. Because the system programs are proprietary and therefore not for sale, this manual will be of practical interest only to federal government agencies which store and maintain time series in the CANSIM data base. Users desiring a fuller knowledge of the codes used and the storage methodology will find this volume a useful adjunct to the "Retrieval Manual".

"CANSIM: Users' Manual for Data Retrieval and Manipulation", Catalogue No. 12-531, price \$1.50. This manual gives detailed instructions on procedures for obtaining data from the CANSIM data base. The initial very rudimentary retrieval package described will be enlarged and improved; the manual will be expanded and modified accordingly. Existing "commands" will, however, continue to be supported.

The DATABANK-MASSAGER System

The DATABANK and MASSAGER programs, which accept input of data retrieved from CANSIM, were first developed by M.C. McCracken at Southern Methodist University in 1964, and later expanded at the Economic Council of Canada. In May of 1966 the Bank of Canada became the first agency other than the Council to make use of the system. During

that summer and fall the National Energy Board and the Department of Finance also began using the system for maintenance and manipulation of the data necessary in their analytical operations. In late 1966 the Dominion Bureau of Statistics accepted the responsibility for the entry of data into the base. Information was stored and maintained by the DATABANK program prior to conversion to the CANSIM direct access memory storage and the Data Entry System. The two programs are now available at a nominal cost from the Bank of Canada.

The DATABANK program is designed to maintain a large number of economic time series on a **single** magnetic tape. Generally, this restricts the number of series that can be handled efficiently to about 10,000. The program allows for the addition, deletion and editing of any series. The data can also be listed, indexed and copied onto other tapes. In other words, the program performs those operations which fall into the general class of file maintenance. The system is designed to work with **any** data which is arranged or arrangeable in a time series format.

The companion MASSAGER program, which carries out the statistical manipulations of the data allocates computer memory and informs the user of the available amount of memory, accepts input from the DATABANK tape, from the CANSIM tape "in MASSAGER format" or from cards, and performs the user specified data manipulations.

Retrieved series are arrayed as columns in core storage and by a sequence of "commands" the columns are manipulated as desired. The commands include simple operations on a single

series (column) such as square roots, logarithms, etc., and complex operations on several variables or columns such as multiple regressions, plots, etc. A partial list of operators is given in Table 2.

TABLE 2. MASSAGER Operation Codes

01 $\log_e x$	17 index	32 rank values
02 $\log_{10} x$	18 collapse	33 three-group values
03 $\sin x$	19 $c+x$	34 instrumental variables regression
04 $\cos x$	20 scaling	35 % change
05 x^w	21 $x+y$	36 weighted moving sum
06 e^x	22 $x-y$	37 output by variable
07 random no. (0,1)	23 $x*y$	38 output by observation
08 dummy (1,0,...)	24 x/y	39
09 time trend	25 move	40
10 constant term	26 squeeze out	41 user-supplied sub-routine XXX1
11 x_t	27 multiple plot	42 user-supplied sub-routine XXX2
12 x_{t-k}	28 plot	43 user-supplied sub-routine XXX3
13 $1/x$	29 multiple regression	44 combined operations
14 cumulator	30 three-pass least squares	46 change location
15 $c*x$	31 nonlinear regression	47 row summation
16 x		

